Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1267 – 1274

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Comparison between Utility Expectation of Public and Private Data in the Market of Data

Teruaki Hayashi*, Yukio Ohsawa

Department of Systems Innovation, School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8656

Abstract

The potential expectation about generating innovative businesses and creating values by combining data from different regions has been increased, however, even the information about data and knowledge of data utilization are not shared. In order to lead data-driven innovations, a market of data is expected to reduce the risks of data utilization, and encourage data exchange. Data Jacket (DJ) has been proposed as the technique for sharing the information about data by publishing the summary of datasets as meta-data. Even if the data itself is not open in public, DJs enable the stakeholders in the market of data to consider the latent value of datasets by understanding the information about data described on DJs. Datasets described in DJs have various sharing policies of data, such as Open Data from governments, or private data from companies or individuals. Although, in general, there is a preconception that "the fee-charging things are better than the free ones", the relationship of sharing policies and the utility expectations of data has not discussed in the market of data. In this study, we examine which data is expected higher for utilizing, that is, comparing the utility expectation of the data which can be shared in public (Public Data) and the data which cannot be shared (Private Data). Observing the frequency of uses in Innovators Marketplace on Data Jackets, which is the gamified workshop for discussing the data utilization, and analyzing the number of views in DJ Store, which is the retrieval system of DJs, the result shows that the data which is hardly open may be recognized valuable for utilizing.

Peer-review under responsibility of KES International

Keywords: Innovators Marketplace; Data Jacket; market of data; scenario

* Corresponding author. Tel.: +81-5841-2908.
E-mail address: teru-h.884@nifty.com

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.171

1. Introduction

In the background of the trend of "big data", it is considered that there is the increase of the potential expectation about generating or improving businesses by combining existent data from different domains, as well as the increase of the stored data. The rapid spread of personal devices enabled service providers to acquire the amount of personal data such as purchase histories or moving records of the consuming public, which had been difficult to be obtained. In the field of industry, the network of physical objects called Internet of Things (IoT)[1] attracts attention, and some companies have taken advantages of the data from other domains or the data in-house, and have started to improve their existent services by adding value. In the field of Open Data, national or local governments have supplied their datasets to the public allowing the secondary use on the Web. Linked Open Data (LOD)[2] has been a recent movement to interlink datasets, based on a concept of Linked Data, i.e., publishing and connecting structured data on the Web.

However, various barriers to data utilization and exchange are pointed out in taking advantage of results of data analyses. Acquiati and Gross mention that the combination of public databases may occur a serious violation of privacy[3]. Xu et al. review the privacy issues related to data mining, by differentiating the responsibilities of different users[4]. The cost of data management and security issues discourage private companies or individuals to open or share their datasets. On the other hand, the problems of combining data are pointed out, as well as the problems of privacy issues. Bollier mentions that combining data from multiple sources does not offer valuable insights, and even it makes the objective interpretation difficult[5]. Boyd and Crawford criticize that the size of the datasets is meaningless without taking into account the sample of each dataset, and suggest the importance of understanding the values of small data stored in different domains[6]. Moreover, accessible datasets in LOD are limited to the data published by governments or research institutions, and the data collected in most companies, individuals, and other research institutions still has been closed. Not only the data itself is inaccessible, but also even the information and knowledge about how to utilize data and make decisions is not available.

Although "big data" has attracted social attention and expectation for data utilization, datasets in private companies, autonomies or individuals are not open or shared due to the cost of data management and security issues. Instead of forcing to open or share data, it is important to activate a market of data[7,8], where users select data, negotiate with data owners, and get data at reasonable conditions, e.g., a price for exchanging datasets. In the market of data, data itself is not necessarily shared, but the information about data as meta-data. Data Jacket (explained later) has been proposed as the technique for sharing the information about data by publishing the summary of datasets. Stakeholders can discuss the methods for data utilization, reliable combination of datasets and analysis tools, and the value of data, using Data Jackets.

Here is a question. What kind of data is valuable for the stakeholders in the market of data? There is a preconception that "the fee-charging things are better than the free ones". Is it correct in the market of data? In other words, do stakeholders in the market of data recognize that the quality of private data, which is obtained by private companies or individuals, is guaranteed, because private data is more difficult to obtain than public data? The relationship of sharing policies and the utility expectations of data in the market of data has not discussed in previous studies.

In this paper, focusing on the sharing policy of data described in Data Jackets (DJs), we discuss the feature and the utility expectation of data, observing the results in the workshops of Innovators Marketplace on Data Jackets, and analyzing users' interests in DJs using the access log of Data Jacket Store, which is the retrieval system of DJs.

The remainder of this paper is organized as follows. In Section 2, we present DJs in detail and discuss the sharing policy of data, looking at the information described in DJs. Section 3 describes experimental details. Section 4 shows the results and the discussion. Finally, Section 5 concludes the paper with a brief discussion.

2. The method for evaluating data: Innovators Marketplace on Data Jackets

2.1. Data Jacket

In order to support stakeholders' discussion in the market of data without publishing data owners' datasets, a Data Jacket[8,9,10] (DJ) has been proposed as the technique for sharing the information about data by publishing the

summary of datasets as meta-data. The feature of DJs is to provide the method for describing datasets in order to consider the potential value of datasets, allowing data itself hidden. DJs consist of structured meta-data represented in natural language, which is recognizable and understandable for humans and computers. Table 1 shows the example of DJ description. We can understand the outline, the format, the data type, variables or the sharing policy of data referring to meta-data on DJs, even if data itself is not open. By publishing DJs, participants of IMDJ can consider the contents of datasets by reading DJs, and start to communicate about data utilizations among data owners, analysts and data users. DJs have been currently registered more than 1,000 (April, 2016)[11].

Table 1. The example of Data Jacket description of "Streetlights Data in Tokyo"

Item	Contents
Title	Streetlights Data in Tokyo
Outline	This data includes the location and the control information of streetlights in Tokyo, which is obtained and supervised by Tokyo Metropolitan Government.
Variable Label	Latitude
	Longitude
	ID number
	Flux of light
	The type of illumination
Sharing Policy	With particular conditions
Data Format	CSV
Data Type	Number
	String

2.2. Innovators Marketplace on Data Jackets

Innovators Marketplace on Data Jackets[8,9] (IMDJ) is a gamified workshop for discussing the data utilization, by providing information about data as DJs. In the process of IMDJ, we introduce tools for data visualization, e.g., KeyGraph[12], and create a map on which shows possible combinations of DJs, which support participants to discover latent combinations of datasets (Fig.1). Data owners provide their datasets as DJs, and participants of IMDJ (including data owners, users, and analysts) create solutions for solving data users' problems stated as requirements. Through the communication among participants, data owners are able to learn how to use their own data from the possible combination of DJs proposed by data analysts, and users are able to learn how their requirements can be satisfied with proposals (we call "solutions" in IMDJ). Through the process of IMDJ, participants who learn the utility expectations of data, start to negotiate for data exchange or buying/selling to create new businesses.

In the communication among stakeholders participating in IMDJ, the potential values of datasets are externalized, and some analysis results were realized. For example, Ikegami et al. report that they achieved to obtain data of streetlights which was the private data, after the negotiation with the local government in Tokyo based on the solution created in IMDJ[13]. The solution is "The application which suggests the safe and secure route in the night by combining the location information of streetlights on the Google Maps", created by combining the streetlight data in Table 1 and Google Maps API. Tangled String[14,15] is one of the tools created based on the solution in IMDJ. Applying Tangled String for the sequential data, such as stock prices or texts of history, Ohsawa et al. suggest that the valuable events may be extracted from the visualization.

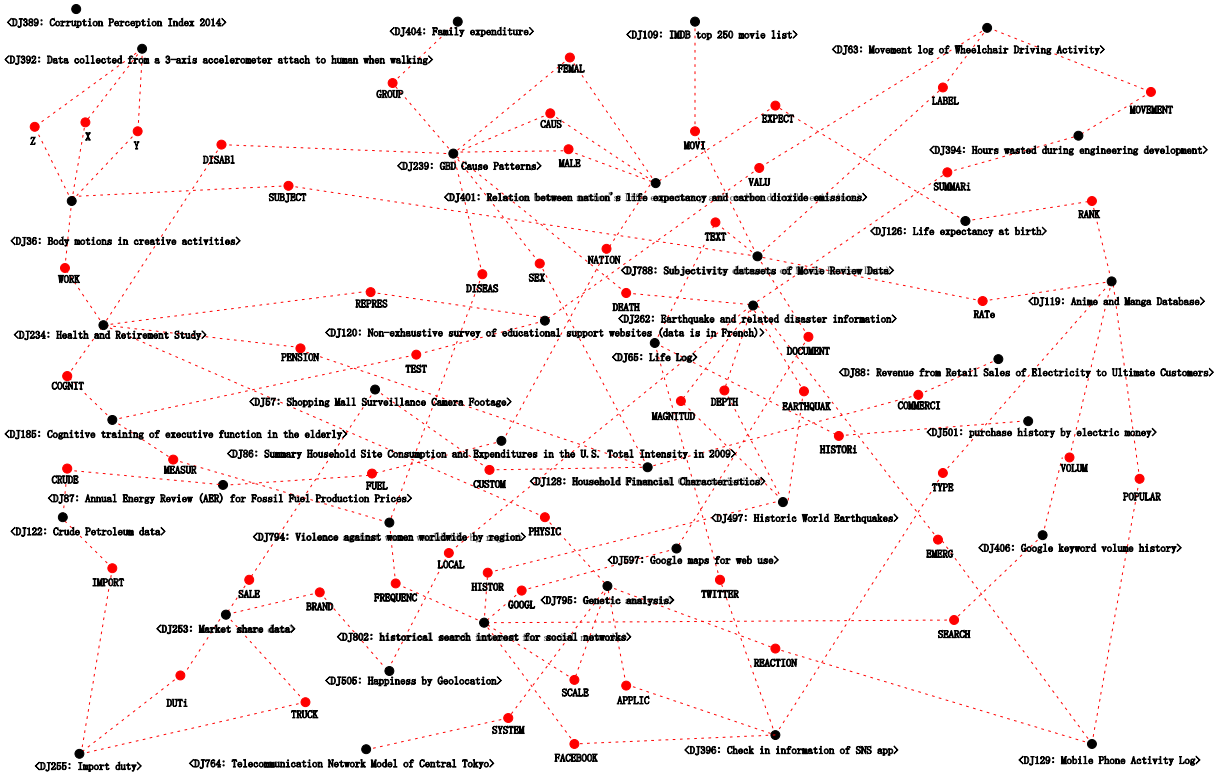

Fig. 1. The example of a visualized map using KeyGraph (black nodes represent DJs, and red nodes shows the keywords included in DJs. Each DJ is interlinked with related Data Jackets via correlated words shown as red nodes, e.g., the words in outlines of data, or in variable labels.)

2.3. The Sharing Policies and the values of data

There are different types of data in the world, which has various features. A sharing policy is one of the important features of data, which means a condition for sharing data. For example, the sharing policy of Open Data is “shared with anyone”, which are the datasets open to the public by some local governments. On the other hand, personal data of medical treatments or customer information of private companies may not be shared, or it may be necessary to negotiate for sharing data. There are some data which can be shared after purchasing, or some data cannot be shared with any other institution. Although there is a variety of sharing policies of data, these can be classified roughly into two groups, that is, the data which can be shared or which cannot be shared. In this paper, we define the data available in public as Public Data, and the data generally unavailable as Private Data. Public Data includes the data already published on the Web or the data which can be disclosed when it is necessary. On the other hand, Private Data is the data which has certain conditions for sharing or the data which cannot be shared considering the risk of disclosure.

The left side of Fig.2 shows the number of DJs divided by their sharing policies, and the right side represents the breakdown of Private Data. The total number of DJs available in DJ Site[11] is 909, and more than half of them (495 DJs) are Public Data. 71 DJs do not have the information about sharing policies of data. This is because the description rule of DJs cannot force data owners to enter all the information about their data.

Looking at the right side of Fig.2, the total number of DJs in the pie chart are 402, because some DJs have several sharing policies, e.g., “data can be purchased only for the research purposes”. 31% (123 DJs) can be shared in particular conditions (including negotiations), and 21% (86 DJs) can be shared partially in a disclosure range. 18% (71 DJs) are impossible to be shared, which are limited only for the in-house use. 14% of data (56 DJs) is undecided, but it cannot be open in public. What is noteworthy in this chart is that 13% of data (52 DJs) are already priced and sold in the market, e.g., the structured data of newspaper articles or meta-data of TV programs.

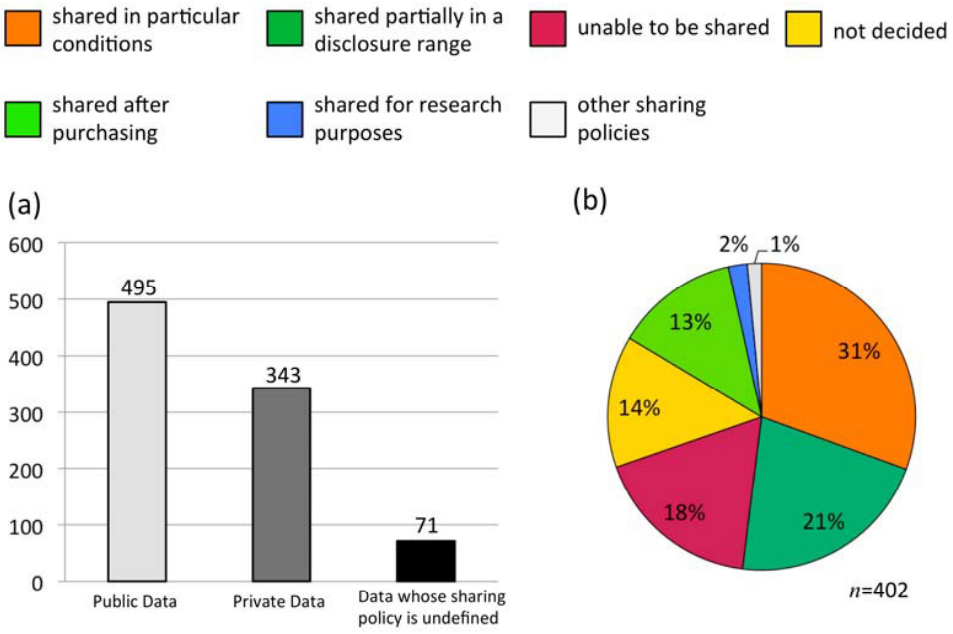

Fig. 2. The number of DJs with their sharing policies (a) and the breakdown of Private Data (b)

As we discussed, the datasets registered as DJs include the information about the sharing policy. In this paper, focusing on the sharing policy of data, we examine the utility expectations of data, by comparing the number of DJs of Public Data and Private Data which users browse in detail, from the access log of Data Jacket Store[16,17] (DJ Store), which is the retrieval system of DJs. Furthermore, in order to compare the utility expectations of Public Data and Private Data in the actual discussion of data utilization, we examine the frequency of uses of Public Data and Private Data in the workshops of IMDJ. The next section presents our experimental study and the experimental steps in detail.

3. Experiment

The purpose of this experiment is to examine which data is recognized valuable, the data which can be shared in public (Public Data) or the data which cannot be shared in public (Private Data). The DJs whose sharing policy is “shared with anyone” are classified in Public Data, and the DJs whose sharing policy has conditions for sharing, e.g., “shared in particular conditions”, “shared partially in a disclosure range”, “unable to be shared”, “not decided”, “shared after purchasing”, or “shared for research purpose”, are classified in Private Data.

In this experiment, in order to examine the utility expectations of data from the users' retrieving activity, we count the number of DJs browsed in detail in DJ Store, using the access log. Comparing the number of Public Data and Private Data, we examine which data are recognized interesting for users. In DJ Store, the list of titles of DJs is obtained when users search with the query in natural language (Fig.3). By clicking on the title of DJs, detailed information, such as variable labels, the outline, or the format of data, is acquired. In this paper, we assume that the browsed DJs in detail are the DJs which users are interested in, and count the number of views separately from sharing policies. Because the purpose of this experiment is to compare the number of DJs focusing on their sharing policies, we do not extract the information about data whose sharing policy is undefined.

909 DJs, 392 requirements and 333 solutions are stored as RDF/XML and reused as data utilization knowledge for the searching system of DJ Store (21,290 triple in total). RDF documents are stored in the SPARQL endpoint, sparqlEPCU[18]. In addition, when users input sentences as queries in DJ Store, extremely high frequent words, such as "data" and "information", symbols such as "." or "," are removed as stop words. We extracted the users' access log obtained from January 31, 2015 to April 16, 2016.

In order to examine the utility expectations of data in the discussions of data utilization, we compare the number of DJs utilized for creating solutions in IMDJ workshops, assuming that the utility expectation of DJs which are frequently used for creating solutions are high. We count the frequency of uses of DJs for creating solutions separately from their sharing policies. We use the dataset of IMDJ workshops stored in DJ Store. In each workshop 10-15 participants joined, and 20-30 DJs were visualized in scenario maps of IMDJ. KeyGraph[12] was used for visualizing DJs. Each workshop was conducted for 1.5 hours.

In addition, the information about the sharing policy of data was not presented to users of DJ Store and the participants of IMDJ workshops.

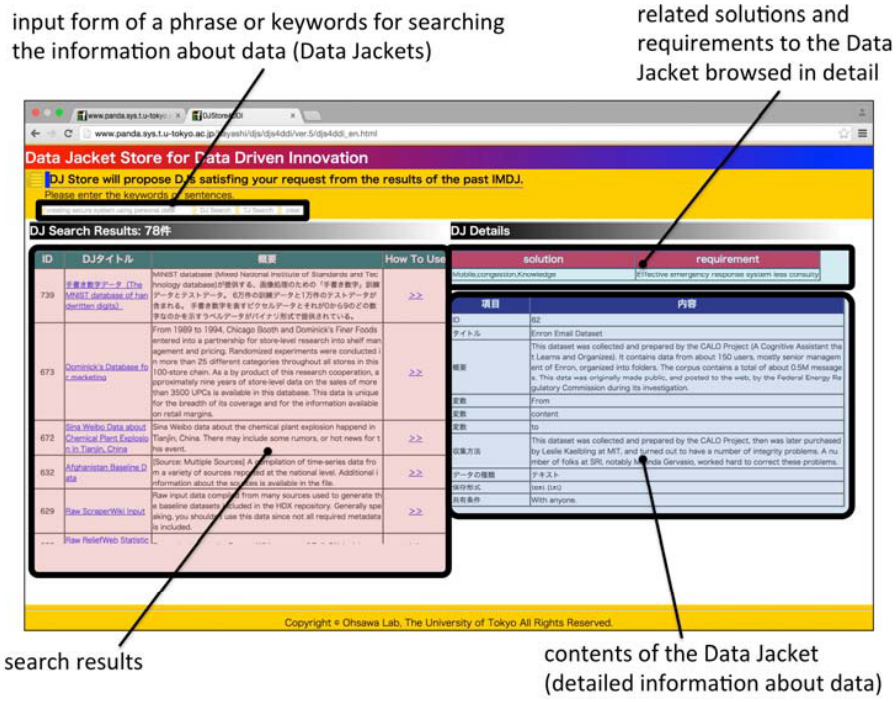

Fig. 3. The snapshot of experimental Web application of DJ Store

4. Result and discussion

4.1. Comparison of the number of views in DJ Store

We obtained the average scores of views from the access log of DJ Store. Table 2 shows the comparison of the number of views between Public Data and Private Data. As explained in Section 2 (Fig.2), because Public Data and Private Data are two independent samples, first, we checked whether these two groups have the same variance. Using f-test, we found that it cannot be assumed that the two distributions have the same variance ($p < 0.01$). With an unpaired t-test assuming the unequal variances, the number of views of Private Data is significantly higher than that of Public Data. This result suggests that users searching the information about data may be interested in data which is not open in public.

4.2. Comparison of the frequency of uses for creating solutions in IMDJ

In comparison of the frequency of uses for creating solutions in IMDJ, we obtained the average scores of uses of DJs in IMDJ. Table 2 shows the result of the frequency of uses. Checking whether Public Data and Private Data have the same variance using f-test, we found that it cannot be assumed that the two distributions have the same variance ($p < 0.01$). With an unpaired t-test assuming the unequal variances, we found that the frequency of uses of Private Data is significantly higher than that of Public Data. This result suggests that participants may recognize that the utility expectations of Private Data are higher than those of Public Data.

Table 2. The results of the number of views and the frequency of uses (average scores ± standard deviation)

Sharing Policy	**The number of views**	**The frequency of uses**
Public Data	1.776±4.470	0.430±1.604
Private Data	4.393±6.366	1.163±2.193
p-value	**	**

**: p<0.01, *: p<0.05, *n. s.*: non significance

With the results shown in Subsection 4.1 and 4.2, it can be said that the data included in Private Data may have higher utility expectations than the data in Public Data. The noteworthy point of this experiment is that the information about sharing policy of data was not presented to the users of DJ Store and the participants of IMDJ. In other words, the users and the participants search or combine DJs tend to choose Private Data unconsciously without knowing the information about sharing policies.

5. Conclusion

In this paper, we discussed the feature of data which is highly expected for utilization in the market of data, focusing on the sharing policy of data. Using the access log of the retrieval system of DJs and the record of discussions of data utilization, we examined the number of views of DJs and the frequency of uses of DJs for creating solutions in IMDJ separately from their sharing policies; Public Data and Private Data. The results of the experiment showed that the number that users browsed Private Data were significantly higher than the number of Public Data. This result suggests that users searching the information about data may be interested in data which is not open in public. Moreover, the frequency that participants of IMDJ created solutions by combining Private Data is significantly higher than the frequency of Public Data. This result suggests that participants may recognize that

the utility expectations of Private Data are higher than those of Public Data in the discussion of data utilization. In conclusion, the results of this paper show that the datasets which are difficult to be shared in public may be expected higher for utilizing, not only when users search the information about data, but also when they think about solutions by combining data.

In this experiment, we divided datasets into Public Data and Private Data, using sharing policies described in DJs. However, there is the difference of views and uses even among Public Data, as well as among Private Data. In the future work, based on the results of this paper, it is necessary to examine the differences between the detailed features of the Public Data and Private Data. Moreover, we found that users and participants tend to choose Private Data, even if the information about sharing policies of data is not presented. It is necessary to discuss the reason why datasets cannot be shared attract the attention of stakeholders in the market of data.

Acknowledgements

This study was partially supported by JST-CREST, and JSPS KAKENHI Grant Number 16J06450. Also we would like to thank to all the staff members of KKE (Kozo Keikaku Engineering Inc.) for supporting our research.

The present research was partially supported through the Leading Graduates Schools Program, "Global Leader Program for Social Design and Management," by the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Manyika J, Chui M, Bisson P, Woetzel J, Dobbs R, Bughin J, and Aharon D, The Internet of Things: Mapping the Value beyond the Hype. McKinsey Global Institute; 2015.
2. Yu L. *A Developer's Guide to the Smantic Web*. Springer; 2011.
3. Acquisti A, and Gross R. Predicting social security numbers from public data, Proceedings of the National Academy of Science, Vol.106, No.27, p.10975–10980, 2009.
4. Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information Security in Big Data: Privacy and Data Mining, IEEE Access, Vol.2, p.1149-1176, IEEE, 2014.
5. Bollier D. *The promise and peril of big data*, Communications and Society Program, The Aspen Institute, Washington, DC; 2010.
6. Boyd D, and Crawford K. Critical Questions for Big Data, *Information, Communication & Society*, Vol.15, No.5, p.662-679, 2012.
7. Liu C, Ohsawa Y, Suda Y. Valuation of Data through Use- Scenarios in Innovators' Marketplace on Data Jackets, ICDMW 2013, 1st Workshop on Market of Data, p.694-701.
8. Ohsawa Y, Kido H, Hayashi T, Liu C, Komoda K. Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data, *Knowledge-based Information Systems in Practice*, Springer-Verlag, 30, p.83-97, 2015.
9. Ohsawa Y, Kido H, Hayashi T, Liu C. Data Jackets for Synthesizing Values in the Market of Data. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 2013; **22**, p.709-716.
10. Ohsawa Y, Liu C, Hayashi T, Kido H. Data Jackets for Externalizing Use Value of Hidden Datasets, 18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 2014; **35**, p.946-953.
11. Data Jacket Site, [Online]. Available from: <https://sites.google.com/site/datajackets/>. [Accessed 21st April 2016].
12. Ohsawa Y, Nels EB, Yachida M. KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, Proc. Advanced Digital Library Conference (IEEE ADL'98); 1998, p.12-18.
13. Ikegami K, Hayashi T, and Ohsawa Y. Creative Communication and Action Process in Utilization of Data (in Japanese), *IEICE Technical Report*, Vol.114, No.343, AI2014-33, p.45-50, 2014.
14. Ohsawa Y, and Hayashi T. Tangled String for Sequence Visualization as Fruit of Ideas in Innovators Marketplace on Data Jackets, *Intelligent Decision Technologies*, p.1-13, 2016.
15. Ohsawa Y. Tangled String Diverted for Evaluating Stock Risks, ICDMW 2015, 3rd Workshop on Market of Data. p.734-741.
16. Hayashi T, Ohsawa Y. Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data, 2nd International Conference on Signal Processing and Integrated Networks; 2015. p.566-571.
17. Data Jacket Store, [Online]. Available from: <http://www.panda.sys.t.u-tokyo.ac.jp/hayashi/djs/djs4ddi/>. [Accessed 21st April 2016].
18. sparqlEPCU, [Online]. Available from: <http://lodcu.cs.chubu.ac.jp/SparqlEPCU/>/ [Accessed 21st April 2016].